Transcrição

Iniciaremos nosso curso de Estatística **na prática**.

O **projeto inicial** pode ser baixado no tópico anterior ou por meio deste link, pois o utilizaremos a partir de agora.

Dentro da pasta "Curso de Estatística", encontraremos dois arquivos: o `Curso_de_Estatística_Parte_1.ipynb` com o notebook que preparamos pa

Com o **Google Colab** aberto, faremos o upload do primeiro acessando "File > Upload notebook...". Na nova janela, acessaremos a opção "Upload" e

Entrando em "View > Collapse sections" ou pressionando as teclas "Ctrl +]" para fecharmos as seções, veremos uma aba lateral com "Table of conte nosso trabalho.

Neste curso básico, veremos os principais tópicos voltados principalmente às **Estatísticas Descritivas**, as quais compõe a primeira fase de um proce

- Conhecendo os Dados;
- Distribuição de Frequências;
- Medidas de Tendência Central;
- Medidas Separatrizes;
- Medidas de Dispersão.

Para começarmos de fato, expandiremos as seções novamente acessando "View > Expand sections" ou pressionando as teclas "Ctrl + [". Com isso, v

O primeiro passo é fazermos o upload do arquivo dataset clicando em "File" na janela lateral. Quando terminar a conexão, clicaremos em "Upload" e

Após isso, um lembrete aparecerá e clicaremos em "Ok"; precisaremos fazer este procedimento de upload toda vez que fecharmos as seções e reinici No passo seguinte, veremos uma maneira de baixar um notebook com as modificações realizadas salvas.

O `dados.csv` é um conjunto vindo do site do Instituto Brasileiro de Geografia e Estatística (**IBGE**). Dentro de "1.1 Dataset do projeto", clicaremos

Nesta página, encontraremos os **microdados**. Adiante, teremos o acesso à pesquisa PNAD 2015 escolhida neste curso, onde cada registro de um mic modificação foi feita a partir de 2015, a qual pode ser encontrada no mesmo portal do IBGE caso seja interessante.

Teremos as informações do questionário, pois é o mais agregado dentro de uma pesquisa estatística.

De volta ao nosso notebook `Curso_de_Estatística_Parte_1.ipynb` , veremos uma descrição básica da Pesquisa Nacional por Amostras de Domicí

Tratamos este dataset utilizando algumas informações disponibilizadas no link de fonte de dados, como a Leitura em **R** que é um software estatístico

Para acompanharmos nossas aulas, usaremos as seguintes **variáveis**:

- A **Renda** calculará o rendimento mensal do trabalho principal para pessoas de 10 ou mais anos de idade;

- **Idade** da moradora ou morador entrevistado contado em anos;

- A **Altura** foi elaborada para o curso, e a entenderemos melhor adiante;

- **UF**: Unidades da Federação com os códigos dos estados brasileiros como um dicionário;

- **Sexo** feminino ou masculino com identificações numéricas também;

- Os **Anos de Estudo** possuem uma codificação de acordo com a quantidade de tempo estabelecida na tabela;

- Por fim, a **Cor** identifica a etnia da pessoa entrevistada com códigos.

Adiante no notebook, encontraremos algumas observações importantes para a área de Estatística e **Ciência de Dados**, pois é bastante interessante do

Fizemos a anotação de que eliminamos registros de renda inválidos, os quais aparecem com o código `(999 999 999 999)` na PNAD 2015.

Também retiramos as `missing` que representam renda nenhuma ou nula, diferente de zero. Por fim, consideramos apenas os registros das Pessoas d

Com isso, poderemos iniciar com a biblioteca **Pandas** para lermos os dados em `.csv` e os passarmos para `DataFrame`. É interessante ter algum co

O primeiro passo é importarmos `pandas` como `pd` com `import` na primeira célula da parte "Importando pandas e lendo o dataset do projeto" do n

```
import pandas as pd
```

Após executarmos as requisições com "Shift + Enter" ou clicando no ícone de "play" da célula, atribuiremos todo o dataset à variável `dados`. Cham
arquivo `dados.csv` entre aspas simples dentro dos parênteses.

```
dados = pd.read_csv('dados.csv')
```

Com isso, o transformaremos em um `DataFrame`. Apenas para vermos seu tipo, escreveremos `type()` recebendo `dados`, e o retorno será `pandas`

Para visualizarmos os dados com o Colab, o qual é muito parecido com o Jupyter Notebook, bastará escrevermos `dados` na célula seguinte e rodarn

Como resultado, veremos um conjunto de dados separados para o curso, contendo as colunas `UF`, `Sexo`, `Idade`, `Cor`, `Anos de Estudo`, `Renda`

O próximo passo será entendermos quando as variáveis são construídas e como as classificaremos. É importante fazermos isso, porque futuramente a
diferentes também.